Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Metodologie e strumenti per lo studio del bias di genere nei documenti di testo

Relatore Prof. Massimo Melucci
Dipartimento di Ingegneria dell'Informazione

Laureando: Matteo Ripamonti
Matricola N 1227738

Anno Accademico 2021/2022

INDICE

1. Introduzione 3

2. Definizioni generali degli argomenti che verranno trattati 6

2.1 Bias e bias di genere 6

2.2 Genere e sesso 7

2.3 *Natural language processing* e *word embedding* 8

2.4 *Word2Vec* 9

3. Dati e metodi 11

3.1 Dati 11

3.2 Metodi 12

3.3 Pulizia e lavorazione dati 14

4. Valutazione modello 16

4.1 Introduzione valutazione 16

4.2 Metodi di valutazione 17

4.3 Descrizione *query dataset* 20

4.4 Ponderazione nella valutazione 22

4.5 Variabili del modello 23

4.6 Risultati e interpretazione 23

4.7 Conclusioni finali di valutazione del modello 27

5. Analisi bias di genere 28

5.1 Introduzioni analisi 29

5.2 Metodologie di analisi 28

5.3 *Query dataset* 30

5.4 Risultati analisi 31

5.5 Conclusioni 32

6. Riflessioni conclusive 34

7. Bibliografia 36

1. Introduzione

I dati digitali sono un artefatto che è al tempo stesso tecnologia e rappresentazione del mondo [1]. Questa è la definizione che il ricercatore e studioso Carlo Batini fornisce nella pubblicazione “Enciclopedia dei dati digitali”.

I dati digitali, di qualsiasi tipologia, forniscono uno specchio, magari non completo ma esaustivo, delle società nella quale viviamo; più dati abbiamo, maggiori informazioni sono disponibili per essere osservate ed elaborate. Viene quasi da pensare che, seppur non siano un bene di prima necessità, li utilizziamo come se lo fossero, attribuendoli un valore molto prezioso.

I dati digitali in sé sono innocui fino a quando non vengono presi e utilizzati con metodi capaci di estrapolare un significato o una rappresentazione di qualche tipo; viene dunque da pensare quanto sia importante trovare degli strumenti che abbiano la capacità di, metaforicamente, trasformare colori in un quadro.

Le diverse tipologie di modelli e tecniche di machine learning sono attualmente dei buoni strumenti che ci danno la possibilità di dare valore ai dati digitali e fornirci una loro rappresentazione. È decisivo però capire in che modo questi strumenti possano diventare dei mezzi utili nella nostra società, senza avere, o solamente minimamente, degli impatti negativi sulle persone o per l’ambiente che ci circonda: in poche parole possiamo parlare di etica dei dati digitali [2].

Uno dei metodi al quale faremo riferimento in questo studio è quello del *word embedding*, ossia uno strumento utilizzato per l’elaborazione del linguaggio naturale tramite l’utilizzo di reti neurali.

Questo metodo risulta ad oggi essere molto usato quando abbiamo a che fare con dati digitali sotto forma di testo. Seppur sia un buono strumento, porsi degli obiettivi per capirne i possibili effetti negativi sotto un punto di vista etico è importante.

Uno tra gli aspetti interessanti negli studi dei dati testuali tramite modelli riguarda i bias di genere, ossia le distorsioni che prendono in considerazione gli stereotipi di genere. Questo tipo di problema è presente in molti ambienti che ci circondano; il fenomeno si manifesta in svariate forme, una delle quali è, appunto, il linguaggio scritto; articoli

giornalistici e pubblicità sono i primi che possono venire in mente. Un altro possibile contesto da prendere in considerazione riguarda la letteratura.

Uno degli obiettivi di questa ricerca sarà elaborare degli strumenti capaci di fornirci informazioni su come la presenza di bias di genere, riguardanti gli stereotipi di genere, siano presenti nel linguaggio di testo, in questo caso romanzi letterali.

Per farlo il lavoro avrà il seguente ordine: nel secondo e terzo capitolo si approfondiranno le definizioni degli argomenti trattati, i dati utilizzati e i metodi principali. Il quarto capitolo è dedicato alla valutazione del nostro modello, cercando di capire quale sia il più funzionale al nostro lavoro; il quinto, invece, riguarda l'analisi dei bias di genere nei romanzi con il modello ottenuto precedentemente, mostrandone i risultati.

La tecnica del *word embedding* verrà impiegata quindi per comprendere la semantica nel linguaggio naturale da parte di un modello. Il funzionamento di base è trasformare le parole in vettori di numeri reali e dopo di ché costruire uno spazio vettoriale in cui i vettori delle parole sono più vicini se le parole occorrono negli stessi contesti linguistici, cioè se sono riconosciute come semanticamente più simili [3].

Infatti, lo strumento del *word embedding* non è l'unico per affrontare i bias e la semantica del linguaggio nel testo, ma è considerato uno strumento di grande utilizzo in questo ambito di ricerca [4].

Utilizzeremo come dati romanzi classici di letteratura, fruibili dal Progetto Gutenberg [5]; più precisamente sceglieremo i romanzi più scaricati, e quindi presumibilmente letti, del portale.

Faremo due assunzioni importanti per il nostro lavoro. La prima è ipotizzare che i dati raccolti, quindi i libri, siano un possibile campione rappresentativo dei romanzi che vengono abitualmente usati e letti dalle persone; generalmente i testi presenti nella raccolta del Progetto Gutenberg fanno riferimento a romanzi che non hanno più diritti d'autore perché appartenenti ad un'epoca passata e non attuale. La seconda assunzione fatta riguarda la correlazione tra scaricare il libro e leggerlo come conseguenza. Questa ipotesi è parzialmente vera se si pensa che un documento di testo in una libreria online gratuita se è il più scaricato allora è il più letto; ma questa affermazione può essere di conseguenza non veritiera a causa del comportamento

delle persone che possono scaricare un documento solamente perché è tra i più popolari e nei primi risultati di ricerca senza poi usufruirne, soprattutto se è gratuito.

È importante sottolineare che questo lavoro non ha il fine di dare una visione generale e strutturata degli stereotipi di genere presente nella letteratura e di come questi possano influenzare le persone, tantomeno di fornire un giudizio sui romanzi che verranno presi in considerazione; non si cercherà di affermare che i bias di genere, forse presenti nei testi utilizzati, siano proporzionalmente generalizzati a tutti l'ambito culturale letterario; tali deduzioni riguardano altri ambiti di ricerca ed affrontabili infatti con strumenti e conoscenze di base molto complesse, talvolta frutto di molteplici lavori.

Qui si affronteranno e analizzeranno solamente alcuni strumenti che possono essere utilizzati per esaminare la presenza di bias di genere in ambito linguistico, si vedranno delle ipotetiche metodologie che possano dare un supporto di comprensione del problema; talvolta questi strumenti saranno limitati al loro uso, possibili perdite di efficienza sono da tenere in considerazione, anche se si cercherà di limitarle, cercando di trovare un buon compromesso con l'efficienza.

2. Definizioni generali degli argomenti che verranno trattati

L'obiettivo di questo capitolo è di offrire una definizione possibilmente esaustiva delle terminologie che verranno usate nel lavoro, affinché si possano ridurre possibili fraintendimenti più avanti.

2.1 Bias e bias di genere

Quando parliamo di bias, in generale, facciamo riferimento alla distorsione di un fenomeno; infatti, il termine viene tradotto letteralmente come 'obliquo' [6].

Nell'ambito dei modelli, nell'informatica e non solo, i bias possono costituire un problema molto rilevante per i modelli che verranno poi applicati negli ambiti sociali, e non solo [7] [8]. Se pensiamo al campo dell'apprendimento automatico (machine learning), tutti i dati che vengono utilizzati per creare modelli da poi utilizzare nella realtà sono intrisi di distorsioni, molte delle volte legati a stereotipi e pregiudizi, andando a rafforzare le varie asimmetrie e discriminazioni sociali [9].

Tutto questo fa sorgere dunque un'ambiguità nei confronti degli algoritmi in alcuni ambiti, ponendo domande relative alle possibili conseguenze sociali, politiche ed etiche.

I bias di genere rientrano in buona parte in queste ultime problematiche descritte; questi tipi di bias sono generati da stereotipi, i quali si verificano in quanto non sono state considerate in modo opportuno le differenze di genere.

La società di oggi, definita di tipo patriarcale, cavalca notevolmente gli stereotipi di genere e di conseguenza fortifica i bias già presenti negli individui. Questa sorta di bias è ovunque, dalle pubblicità in internet che si rivolgono maggiormente ad un pubblico maschile se devono vendere un'automobile e ad un pubblico femminile per i vestiti [10], le raccomandazioni musicali nei vari servizi streaming offrono certe categorie musicali ad un pubblico femminile piuttosto che maschile [11], i servizi di reclutamento per il lavoro tendono a proporre mansioni in base al genere dell'individuo [12] e così via. Questi esempi fanno riferimento in particolare a modelli che, in base ad input iniziali di dati, sono influenzati nel momento in cui devono prendere decisioni in modo autonomo e si comportano di conseguenza con gli stessi meccanismi.

Un altro contesto, sempre importante, nel quale sono presenti bias di genere è il linguaggio scritto e parlato; romanzi, giornali, dialoghi di film, poesie, tweet e tutte le possibili forme in questo ambito, sono un ambiente favorevole alla diffusione degli stereotipi di genere. I documenti in formato testuale sono uno strumento efficacie per la comunicazione e la proiezione delle interazioni umane, costituiscono l'intero complesso del mondo sociale.

Ed è per questo che gli strumenti e i metodi che verranno sviluppati in questo lavoro prenderanno in considerazione come dati i romanzi letterali, ossia una delle forme di comunicazione indiretta più utilizzate al giorno d'oggi.

Nel nostro lavoro si definirà la presenza di un bias di genere quando un termine, che sia un aggettivo, un nome o un verbo, è maggiormente vicino a pronomi o nomi di un genere rispetto all'altro, ossia che certe parole vengano usate in modo differente dipendentemente al genere di riferimento. L'idea di equità invece viene rappresentata se, indipendentemente dal genere, le parole si distribuiscono in modo uniforme.

2.2 Genere e sesso

Un importante appunto riguarda la distinzione tra genere e sesso, talvolta identificati con lo stesso significato.

Per definizione il sesso è Il complesso dei caratteri anatomici, morfologici, fisiologici (e negli organismi umani anche psicologici) che determinano e distinguono tra gli individui di una stessa specie, animale o vegetale, i maschi dalle femmine e viceversa [13]. Quando parliamo di genere invece si fa riferimento alla distinzione in termini di appartenenza all'uno o all'altro sesso, non in quanto basata sulle differenze di natura biologica o fisica ma su componenti di natura sociale, culturale, comportamentale. [14]

Nelle analisi di dati testuali è possibile ricorrere ad errori che vadano a confondere queste due diverse definizioni; se si fa riferimento, ad esempio, agli organi riproduttivi, si sa bene che certi termini verranno quasi sempre accorpati al sesso di riferimento, quindi in questo caso le differenze di genere non entreranno in gioco; invece se si parla, ad esempio, di una descrizione sotto il punto di vista emotivo, sarà il genere ad essere protagonista di osservazione, perché confuso per la maggior parte delle volte che in base alla propria identità allora ci sia una caratteristica di tipo emotiva specifica.

Ovviamente comprendere a pieno la distinzione e il significato di questi due termini è assai complesso, talvolta fatto in modo erroneo; in questo lavoro si farà riferimento al termine genere per come lo si è appena descritto. Tradizionalmente il sesso e il genere viaggiano in maniera abbastanza legata, ossia che uno determini l'altro, per questioni culturali e sociali.

Un'altra osservazione tanto utile che guarda al genere è la seguente. In questo lavoro si ragionerà con l'idea che l'identità di genere sia una variabile dicotomica, ossia o femmina o maschio. Sottolineiamo che questa è solamente un'approssimazione e che sono presenti una moltitudine di forme che riguardano l'identità di genere, non riconducibili a categorie specifiche, ma che, essendo componenti di natura sociale, culturale e comportamentale, non sono facilmente osservabili e analizzabili.

2.3 *Natural language processing* e *word embedding*

Il *Natural Language Processing* (NLP) è un sistema per il trattamento automatico delle informazioni scritte o parlate del linguaggio naturale; infatti, la comprensione del linguaggio da parte di una macchina risulta essere utile ma complessa. Si può ad esempio volere interpretare un dialogo tra due persone e cercare di capire di cosa stiano parlando, oppure creare un sistema di controllo ortografico automatico. A differenza dei linguaggi di programmazione, il linguaggio umano non è sempre facilmente rappresentabile, presenta forme e strutture che si differenziano in base al luogo, al tempo, il tipo di testo che può essere un discorso, una poesia o una canzone, ecc., per questo è possibile che gli strumenti che cercano di darne una comprensione automatica possono risultare limitati.

Come già accennato precedentemente, uno dei sistemi che utilizzeremo per il NLP è il *word embedding*, ossia un insieme di strumenti e tecniche che permettono di rappresentare parole e frasi di testi scritti attraverso vettori di numeri reali nella seguente maniera:

$$V \rightarrow \mathbb{R}^D : w \rightarrow \vec{w} \qquad (2.1)$$

La mappatura avviene partendo da una parola w appartenente al dizionario V, trasformandola in un vettore a valori reali $\vec{w}$ in uno spazio di dimensione D.

La rappresentazione viene definita in base all'uso delle parole. Ciò consente alle parole che vengono utilizzate in modi simili di avere rappresentazioni simili, catturandone in qualche modo il significato. Quindi per l'analisi del testo si vanno ad analizzare i vettori presenti nello spazio vettoriale attraverso la loro posizione, vicinanza e distanza, da altre parole e così via.

Il vantaggio del *word embedding* è la possibilità di trasformare un corpo testuale in uno vettoriale ed avere quindi la possibilità di lavorare con delle matrici; quindi, non trasforma solamente dei termini semanticamente vicini in punti vicini nello spazio, ma traduce concetti anche più complessi creando una sorta di geometria spaziale del significato.

Provando a pensare ad un esempio tanto banale ma utile che viene utilizzato nella letteratura scientifica del NLP, valutiamo il seguente esempio. [Fig. 2.1]

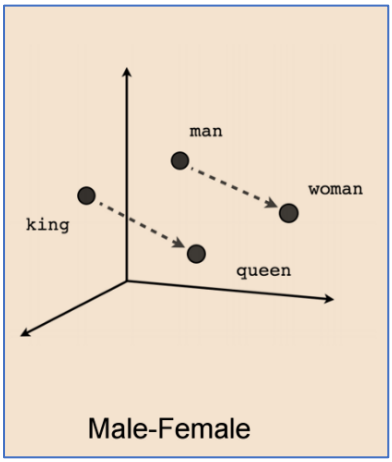

Figura 2.1

In questo caso il nostro spazio vettoriale ha tre dimensioni, i termini sono rappresentati tramite vettori e sono distribuiti in una forma che può fare intendere la vicinanza di quest'ultimi. Il genere femminile e maschile può essere quindi utilizzato per capire quali termini sono maggiormente vicini ad altri per la loro posizione.

2.4 *Word2Vec*

Esistono svariati algoritmi che permettono il *word embedding*, uno dei quali è *Word2Vec*, sviluppato da Tomas Mikolov; grazie alla sua efficienza in questo ambito

[15], il suo utilizzo ha avuto una forte attenzione nell'ambito del NLP, mostrando una buona capacità nel fornire una rappresentazione vettoriale delle parole.

Più precisamente sono state fornite due differenti architetture all'interno di W2V: CBOW (continuous bag-of-words) e Skip-gram, entrambe con una buona efficienza [Fig. 2.2]. Si ritiene però che, sempre secondo Mikolov, l'architettura Skip-gram lavora meglio con una minore quantità di dati e rappresenta meglio termini più rari; a sua volta CBOW è più veloce e lavora meglio con termini che si ripetono maggiormente. Entrambi i modelli sono implementati tramite una rete neurale artificiale costruita su tre strati: uno strato di ingresso (*input layer*), uno strato nascosto (*hidden layer*) e uno strato d'uscita (*ouput layer*).

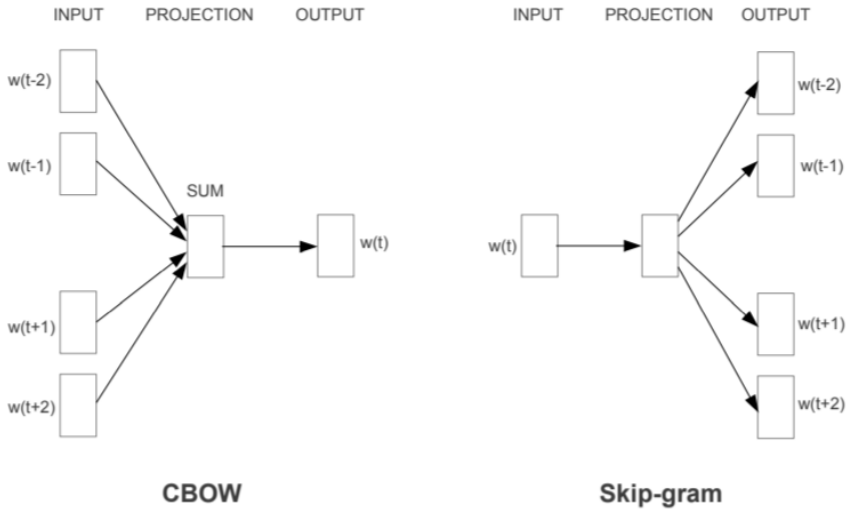

Figura 2.2

La differenza più generale sta che nel modello CBOW si cerca di prevedere una parola sulla base dei termini circostanti, mentre Skip Gram cerca di predire i termini circostanti di una parola.

Il pacchetto per l'implementazione in Python è fornito da Radim Rehurek [16] attraverso la libreria Gensim.

3. Dati e metodi

3.1 Dati

Come già accennato in precedenza, verranno presi in considerano come dati dei romanzi letterali; più precisamente l'intero corpo comprenderà circa seicento libri, tutti scritti in lingua inglese.

I libri sono stati resi disponibili grazie al Progetto Gutenberg, ossia una libreria digitale online contenente più di 60mila testi, usufruibili gratuitamente. Essendo un progetto senza scopo di lucro che mette a disposizione testi in modo completamente gratuito, sono presenti per lo più romanzi relativi ai grandi classici letterali del passato.

Per cercare di rimanere coerenti all'obiettivo e ottenere risultati con una certa corrispondenza alla realtà, verranno presi i primi seicento testi più popolari, ossia i più scaricati dal sito. È giusto sottolineare che ogni libro avrà lo stesso peso nella valutazione di tutti gli altri, indipendentemente da quante volte è stato scaricato.

L'utilizzo di libri come dati è solo una tra le tante possibilità di scelta; database relativi a Twitter, dialoghi di una serie televisiva, l'intero corpo delle biografie di Wikipedia e così via. Le risorse relative a possibili lavori inerenti ai bias di genere nel linguaggio sono illimitate.

Inoltre, effettuare un lavoro di NLP tramite *word embedding* per i romanzi testuali, può portare alcuni rischi.

Prima di tutto il linguaggio che viene utilizzato varia in base al luogo, al periodo e al tempo nel quale viene prodotto; le forme nell'ambito della linguistica sono tanto varie quanto complesse che già un possibile studio sotto un punto di vista qualitativo richiede particolari conoscenze; l'interpretazione di una poesia, quindi una forma di linguaggio con strutture poco standard, e suoi possibili studi relativi ai bias di genere, risulta essere già complessa per un essere umano, richiedendo conoscenze relative alla sintassi e alla semantica utilizzata; pensare di affidare un compito così articolato e strutturalmente difficile ad una macchina e ottenere risultati di un certo valore è al momento complesso. Piuttosto ci si può aspettare di costruire strumenti che possono fornire un appoggio parziale ma in qualche modo efficacie a questi ambiti, senza pretendere di ottenere il modello perfetto.

Sarebbero ancora tante le cose da dire riguardo all'uso del linguaggio scritto come dati, ma non è questo il luogo e il tempo di farlo, le criticità sono molteplici e la loro descrizione non risulterà sufficientemente esaustiva.

3.2 Metodi

Il metodo principale che verrà utilizzato, applicato dal *word embedding*, deriva dalla semantica distribuzionale [17][18]; l'idea principale si basa sull'ipotesi che i termini che tendono a ricorrere in contesti linguistici simili siano anch'essi simili, anche semanticamente.

Questo permette di vedere il testo sotto esaminazione come una struttura lessicale ben definita, dove la distribuzione delle parole nello spazio testuale ricopra un ruolo chiave per capirne il loro significato. Ogni termine è presente una o più volte in un testo, affiancato da altri termini, e grazie a quest'ultimi è possibile risalire al suo possibile significato che assume in quel determinato contesto linguistico, senza per forza conoscerne direttamente la sua definizione generale. Secondo questa logica, la compagnia che frequenta abitualmente una parola un testo, va definire chi è quest'ultima. Ovviamente non è detto che questo metodo possa portare ad estrarre sempre il vero significato del termine di riferimento; come già accennato nel punto 2.3, le forme testuali a volte sono più complesse in termini di struttura e non seguono una forma standard come può essere quella discorsiva.

Oltre alla questione della struttura testuale, un altro aspetto riguarda la qualità dei dati testuale, essendo loro il punto di riferimento dal quale partire per estrarre informazioni; se quest'ultimi risultano troppo approssimativi, inesatti e/o intrisi di stereotipi, si potrebbero ottenere degli esiti altrettanto erronei e poco veritieri.

Le peculiarità della semantica distribuzionale rimangono comunque molto interessanti, perché affrontano ambiti legati alla linguistica computazionale e alle scienze cognitive, di conseguenza utili per lo studio dei bias nei documenti di testo.

In seguito, un esempio molto ricorrente è fornito dal linguista computazionale Alessandro Cenci.

Si descrive uno spazio semantico di parole con una quadrupla di valori: { T, B, M, S }

- T è l'insieme delle parole target.

- B è la base che definisce le dimensioni dello spazio geometrico e contiene i contesti linguistici.
- M è una matrice di co-occorrenza che fornisce una rappresentazione vettoriale di ogni parola target T.
- S indica la metrica impiegata per misurare la similarità cioè la distanza tra i punti nello spazio.

Ora si deve rappresentare ogni parola come un vettore a *n* dimensioni, ciascuna delle quali conterà la frequenza con cui la parola appare in un determinato contesto linguistico definito dalla base B. Ogni parola target T corrisponderà a una riga della matrice M e le cui colonne invece agli elementi nella base B.

Si inizia contando in un ipotetico corpus quante volte i nomi auto, gatto, cane e camion co-occorrono con i verbi mangiare, guidare e correre, ottenendo così la seguente matrice di frequenza:

	Mangiare	**Guidare**	**Correre**
Auto	0	3	2
Gatto	4	0	3
Cane	3	0	4
Camion	0	2	3

Per costruire i 4 vettori di auto, gatto, cane e camion si utilizzerà come primo componente la frequenza di co-occorrenza con mangiare, come secondo guidare e infine come terzo correre.

v1 = auto = (0, 3, 2)

v2 = gatto = (4, 0, 3)

v3 = cane = (3, 0, 4)

v4 = camion = (0, 2, 3)

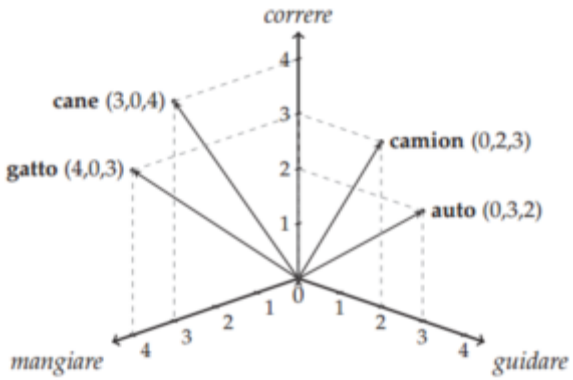

Figura 3.1

Si è quindi fissato uno spazio tridimensionale di concetti e distribuito i vettori delle parole; lo spazio che andranno ad occupare li porterà ad avere posizioni simili per concetti ritenuti simili.

Si vede che dall'esempio lo spazio tridimensionale porta ad avere le parole 'gatto' e 'cane' più vicine, perché con caratteristiche più simili. La stessa cosa vale gli altri due termini [Fig. 3.1].

Il metodo della semantica distribuzionale che viene utilizzato nel *word embedding* porta ad una generale forma di comprensione da parte della macchina in maniera autonoma.

3.3 Pulizia e lavorazione dati

Un altro aspetto importante in questi casi riguarda la pulizia dei dati e la loro lavorazione. L'utilizzo del *word embedding* preclude che ci sia una struttura dati ben definita, non essendo sufficiente utilizzare i dati testuali in forma grezza. Il corpus da utilizzare è costituito da una lista di frasi dell'intero dataset suddivise in base alla punteggiatura di inizio e fine discorso. Queste verranno poi utilizzate nella fase di creazione dello spazio vettoriale delle parole.

Un altro passaggio riguarda la fase di text normalization, ossia effettuare una lavorazione del testo tramite possibili metodi, al fine di renderlo maggiormente funzionale nella fase di *word embedding*. In questo caso è stata effettuata la rimozione delle *stop words* e tutti i caratteri sono stati ridotti in forma minuscola. La parte dell'esclusione delle *stop word* permette di rimuovere tutti quei termini che non sono necessari ai fini del processo, come le preposizioni, articoli e congiunzioni, non avendo

un valore semantico nel testo; in questo modo si presuppone che ci sia una maggiore efficacia ed efficienza dei risultati, riducendo i tempi di elaborazione e dello spazio di memoria e migliorandone la comprensione semantica del testo.

Esistono anche altri metodi di pulizia di dati, come lo *stemming* e il *lemmatization*, i quali hanno sempre lo scopo di ridurre la complessità del testo e mantenere le informazioni più rilevanti a fini del processo. In questo caso però si è deciso di utilizzare gli strumenti descritti precedentemente, perché ritenuti sufficienti.

Concludendo, la fase di *text normalization* permetto sì di semplificare i dati a disposizione ed avere le informazioni più essenziali, ma di contro si può commettere l'errore di semplificare troppo il testo sotto osservazione e perdere importanti riferimenti durante la comprensione della semantica da parte del modello.

4. Valutazione modello

4.1 Introduzione valutazione

La valutazione di un modello ricopre un aspetto fondamentale per la comprensione dei dati; in questi casi il termine valutazione viene utilizzato per indicare il giudizio di uno strumento, il quale possa essere usato per studiare e individuare certe caratteristiche di un fenomeno; è necessario studiarne i suoi pregi e le sue criticità, onde evitando di cadere nella trappola di creare un modello che si adatti solamente ai dati del proprio studio e forzarlo a trovare e ottenere i risultati desiderati. È importante comprendere che valutare un modello è parte sostanziale di questo lavoro, dato che il nostro obiettivo è ottenere dei metodi e degli strumenti che vadano ad individuare i bias di genere.

In questo capitolo verranno proposti dei metodi di valutazione del modello e i risultati ottenuti, con lo scopo di identificare quali caratteristiche sono necessarie al fine di costruire degli strumenti efficaci per il nostro studio. I dati utilizzati saranno i romanzi letterali e dei query dataset, i quali contengono i giudizi relativi ad alcune coppie di parole; successivamente verranno messi a confronto quest'ultimi con i giudizi ottenuti dal nostro modello riferito ai romanzi.

Esistono differenti modelli che studiano il *word embedding*, ognuno con le sue caratteristiche e i suoi parametri di riferimento; in questo studio si utilizzerà come riferimento al modello CBOW (*continuous bag-of-words*) sviluppato da Tomas Mikolov. [15]

Al fine di capire quale modello sarà il più adatto al nostro lavoro, valuteremo l'architettura CBOW per la quale verranno modificati i parametri relativi a: dimensione del vettore-parola, finestra di contesto, ossia decidere la massima distanza tra la parola corrente e quella predetta all'interno di una frase, e dimensione del nostro dataset di dati; verranno dunque utilizzate queste tre variabili. All'interno dello strumento *Word2vec* sono presenti ulteriori parametri da utilizzare per la generazione del modello, ma si è deciso di mantenere i valori di default assegnati dagli autori.

Bisogna però fare un appunto con cosa si intende per dimensione dei dati. Nel lavoro l'unità di riferimento è il quantitativo di libri, quindi, non abbiamo realmente una misura statica unitaria visto che la dimensione di un testo difficilmente è la stessa. Una

possibile unità di misura più precisa e meno generalizzata potrebbe essere quella di tenere conto il numero totale di parole di tutto il dataset. Facendo però una analisi esplorativa iniziale, è stato osservato che mediamente il numero di parole per ogni libro, rispetto ai seicento disponibili per il nostro lavoro, tende a divergere ad un quantitativo simile dopo già venticinque testi; l'unità quindi adottata generalizza questo aspetto per rendere più semplice la comprensione, senza però perdere efficacia.

Il rischio che si corre nell'effettuare una valutazione di questo tipo è, come accennato prima, di creare degli strumenti di giudizio che si adattino solamente ai dati in possesso e non all'obiettivo che si sta affrontando, finendo per costruire uno strumento fine a sé stesso senza un'effettiva validità. Per evitare tale problema si cercheranno di utilizzare metodi che avranno una significatività a livello logico linguistico e statistico.

4.2 Metodi di valutazione

Affinché si possa valutare un modello è necessario trovare dei metodi che siano di supporto in questa fase. In letteratura è presente una buona dose di possibili strumenti con tale scopo, ma non esiste uno strumento specifico standardizzato per ogni situazione sempre valido; dunque, è importante capire quali siano le proprie esigenze di valutazione, individuare quali siano le caratteristiche del nostro modello che si vogliono mettere maggiormente in luce e a quali scopi siano indirizzate [19].

Uno metodo ampiamente utilizzato è attraverso la valutazione intrinseca [20][21], ossia comparare il giudizio del nostro modello verso il giudizio umano nelle relazioni tra parole, ad esempio, valutando la similarità tra due parole; in poche parole, un essere umano fornisce un giudizio sotto forma di punteggio che valuta quanto simili sono due termini, che quindi sarà ritenuta come verità, e la nostra macchina fornirà un giudizio anch'essa, sempre sotto forma di score: l'obiettivo è avere un modello che fornisca dei giudizi il più possibili fedeli alla logica umana.

Alcune criticità che posso essere presenti in questo metodo riguardando la soggettività dei giudizi dell'essere umano. Quando si costruiscono dei dataset riguardanti il significato e/o la comprensione di uno o più termini fornita tramite un giudizio, si deve tenere conto che quest'ultimo non rappresenta oggettivamente la realtà, anche perché nel linguaggio il concetto di comprensione è assai complesso e non unico, ma generalizzato a qualche forma che viene maggiormente utilizzata nella società, o almeno in una parte. Standardizzare la comprensione è assai rischioso, può causare

distorsioni; per ironia della sorte se si volesse valutare la presenza di un bias in un testo e la persona che lo sta facendo utilizza strumenti a sua volta intrisi di ulteriori bias, rischia di rappresentare quella valutazione in modo distorto.

Ci si chiede quindi come si potrebbe affrontare e, se non risolvere perlomeno correggere, questo problema. La risposta più banale cade nel cercare maggiore eterogeneità nei giudizi dati dalla persona; questo non vuol dire ricoprire tutti i possibili giudizi di ogni individuo, ma valutare l'esistenza di possibili gruppi di giudizio differenti.

I metodi che verranno utilizzati sono i seguenti: *Word Similarity e Word Analogy.*

A. Word Similarity: L'idea è quella di valutare la somiglianza tra due termini attraverso un punteggio, come descritto nell'esempio precedente; il modello per dare un punteggio di giudizio misurerà la similarità attraverso il coseno dei vettori riferiti ai due termini messi a confronto, che quindi sarà compreso tra -1, se le due parole sono 'distanti', e 1, se i termini sono vicini tra loro. Per comodità di interpretazione grafica si effettuerà una trasformazione per avere un risultato compreso tra [0, 1].

La seguente formula mostra come avviene il calcolo: le parole messe a confronto sono indicate tramite i rispettivi vettori $\vec{a}\ e\ \overrightarrow{b}$.

$$\cos\theta = \frac{\vec{a} * \vec{b}}{\|\vec{a}\| * \|\vec{b}\|} = \frac{\sum_1^n a_i * b_i}{\sqrt{\sum_1^n a_i^2} * \sqrt{\sum_1^n b_i^2}} \quad (4.1)$$

Dopodiché questo punteggio verrà messo a confronto con il punteggio fornito dal giudizio di un essere umano e si misurerà il grado di correlazione complessivo dei termini utilizzati.

Questo metodo ricopre degli svantaggi da non sottovalutare [22]. La nozione di similarità semantica (*similarity*) tra due termini è spesso soggettiva e a volte confusa con la nozione di parentela semantica (*relatedness*); la prima quantifica la somiglianza di due concetti in base alla loro posizione all'interno di una categoria (ex. auto-treno), la seconda invece quantifica il grado con cui due parole sono associate tra loro (ex. auto-strada): insomma, c'è un aspetto di tipo linguistico che è difficile colmare. Per risolvere questo problema sarebbe necessario costruire dei dataset di giudizi distinti in

base alle due tipologie. Per semplicità si utilizzerà la parola similarità in senso generalizzato. Un altro esempio è la difficoltà a individuare e valutare nel modo corretto le parole polisemantiche, ossia un termine scritto allo stesso modo ma con significato differente avrà la stessa rappresentazione vettoriale e lo stesso giudizio.

Infine, per confrontare e valutare i giudizi del modello e dell'essere umano verrà utilizzato il coefficiente di correlazione di Pearson, per il quale si effettuerà anche un test di ipotesi per verificare se sia presente assenza di correlazione significativa tra i giudizi.

Per sostenere che il nostro modello fornisca una buona interpretazione della somiglianza tra termini, e quindi si trovi un buon coefficiente di correlazione, si è deciso di osservare quelli trovati nei lavori [20] e [21], nei quali vengono confrontati differenti possibili metodi di *word embedding*, ma con query dataset identici a quelli utilizzati in questo lavoro. Teniamo anche conto che la capacità di comprensione del modello in quei lavori è basato su dataset molto più grandi e già dotati di una buona validità, quindi teoricamente affidabili.

B. ***Word Analogy**:* Questo tipo di analisi ha l'obiettivo di verificare l'analogia tra termini. Più precisamente si sostiene che data una coppia di parole X e X' e una terza Y, la relazione di analogia tra X e X' consente di trovare la corrispondente parola Y' a Y; la forma è la seguente:

$$X : X' = Y : Y' \quad (4.2)$$

Un esempio molto usato nella letteratura è:

$$Man : King = Woman : Queen \quad (4.3)$$

Data la coppia di parole *Man* e *King* e il termine *Woman*, il nostro modello dovrebbe assegnare il termine *Queen* per verificarne l'analogia; è da sottolineare che la logica

del nostro modello lavora sotto vettori numerici di parole, di conseguenza ci si aspetterà in termini vettoriali la seguente equazione:

$$\overrightarrow{Queen} = \overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} \quad (4.4)$$

Anche in questo caso l'analogia che corrisponde alla realtà è stabilita da un essere umano, il nostro modello dovrà cercare di ottenere gli stessi risultati.

Effettuando questa operazione per tante forme di analogie, verrà calcolata la proporzione di risposte corrette sul totale di analogie presenti.

Per capire se la valutazione sia andata a buon fine, come fatto prima, si andrà a paragonare il risultato con i lavori [20] e [21].

4.3 Descrizione *query dataset*

Per effettuare una valutazione consistente che riguarda i punti descritti in precedenza, bisogna ricorrere a dataset di giudizi umani (*query*) relativi alle analogie e le similitudini tra parole. È importante prendere in considerazione delle *query* che abbiano una validità e una credibilità buona; infatti, visto che la creazione di questi giudizi è stata fatta manualmente secondo la logica di individui, decidere quale sia affidabile e quale sia più adatto alle nostre circostanze è importante.

Dovendo affrontare una valutazione del modello che andrà poi a lavorare sui bias di genere, sarebbe logico cercare dei dataset di *query* che pongano un'attenzione, parzialmente o totalmente, sull'uso, ad esempio, dei pronomi di genere. Valutare un modello che abbia la capacità di individuare quale siano le forme di analogie e similitudine riguardante, ad esempio, le capitali e le nazionalità, non sarebbe funzionale al nostro lavoro.

I *query dataset* utilizzati per le similarità tra termini sono nove [Fig. 4.1].

Nome	Coppie	Anno
MC-30 [23]	30	1991
MEN [24]	3000	2012
Mturk-287 [25]	287	2011
Mturk-771 [26]	771	2012
RG-65 [27]	65	1965
RW [28]	2034	2013
WS-353-SIM [29]	203	2009
WS-353-REL [29]	252	2009
YP-130 [30]	130	2006

Figura 4.1

La tabella soprastante presenta il nome del *query dataset*, il numero di coppie di parole con il grado di similarità e l'anno della loro creazione.

Ognuno di questi dataset è strutturato nella forma: termine 1, termine 2, grado di similarità.

Per quanto concerne la valutazione per le analogie tra parole, si è optato per il dataset di Google [5], contenente una decina di migliaia di analogie. Sono suddivise in due categorie, semantiche e morfo sintattiche. Per rendere il dataset più affidabile si è deciso di rimuovere l'intera parte riguardanti le analogie tra città e nazionalità, perché ritenute di poco interesse per il tipo di valutazione attuale.

Questi dataset sono protagonisti di criticità. La maggiore parte di loro fa riferimento ad un linguaggio di tipo moderno e valuta i concetti sotto un punto di vista relativo al presente. Utilizzando dei testi che sono stati scritti nel passato, è possibile che i significati di alcuni termini sabbiano delle connotazioni differenti, di conseguenza la qualità delle *query* potrebbe perdere di efficacia. Idealmente bisognerebbe creare dei dataset di interrogazione ad hoc per il nostro lavoro, ma questo richiederebbe tempo e conoscenze e capacità al di fuori di questo studio.

4.4 Ponderazione nella valutazione

Per risolvere, o almeno correggere il problema dei *query dataset* che sono poco adattabili al nostro lavoro descritto precedentemente, si pondererà il peso di una *query* in base al suo valore nei nostri dati.

Sottolineiamo che oltre ad un problema di qualità dei termini presenti nel *query dataset*, ossia che con il tempo il loro significato potrebbe avere subito variazione, è presente anche un problema di frequenza dei termini. Banalmente nei romanzi di testo utilizzati nel nostro lavoro contengono presumibilmente termini che oggi vengono poco utilizzati, di conseguenza vale lo stesso discorso per il dataset di query: è possibile che siano presenti parole non ancora esistenti nel passato o comunque poco utilizzate.

Per quanto riguarda termini presenti nelle *query* ma non nei nostri dati, non potendone dare una valutazione, non sono stati presi in considerazione. Però, sotto il punto di vista della frequenza è possibile effettuare un lavoro di ponderazione, ossia dare più peso ai termini che compaiono maggiormente nel nostro dataset.

Sono state usate metodologie simili sia per *word similarity* e *word analogies*, ossia abbiamo trovato i quantili nella libreria delle frequenze del nostro dataset e successivamente in base al quantile di appartenenza dei termini presenti nella query si dava più peso a quella stessa *query*.

Per le similarità è stata trovato il quantile mediana; se il termine della query fosse stato presente nel primo sarebbe stato assegnato punteggio 0 e se nel secondo quantile 1. Essendo due i termini per ogni query, il punteggio era compreso tra [0,2].

Nelle analogie tra parole, quindi struttura composta da quattro termini, sono stati trovati i quartili; dopodiché veniva assegnato un punteggio di 0 se il termine faceva parte del primo quartile, 0.25 del secondo, 0.75 per il terzo e 1 per il quarto. Il punteggio era compreso nell'intervallo [0,4]. Se il punteggio fosse risultato essere non intero si sarebbe approssimato per difetto.

In breve, se una coppia a confronto avesse ottenuto un punteggio 2, la valutazione di quella coppia verrà presa in considerazione due volte. Se il punteggio è pari a 0, allora, ritendendo una frequenza troppa bassa dei termini, la valutazione non verrà fatta.

L’obiettivo di questo metodo non è quello di aumentare o diminuire la valutazione del modello, ma cercare di migliorare la qualità dei query dataset applicati al nostro dataset.

4.5 Variabili del modello

Come spiegato inizialmente, sono stati valutati diversi modelli in base a tre variabili differenti:

- Dimensione dei dati: Riguarda il numero di libri presi in considerazione divisa in sei categorie (10, 25, 50, 100, 300 e 600 libri).
- Spazio vettoriale: Indica la dimensione vettoriale di ogni parola trasformata in vettore, suddivisa in quattro possibili categorie. Le differenti dimensioni vettoriali sono 50, 100, 300 e 600.
- Dimensione finestra: Ossia la finestra contesto. Spiega la distanza massima dei termini presi in considerazione rispetto alla parola obiettivo durante la fase di *word embedding* per coglierne la semantica. In questo caso la variabile possiede cinque possibili categorie (1, 2, 5, 7 e 10).

I modelli sono stati valutati in base a queste tre variabili, prese una alla volta senza interazioni tra di esse. Ad esempio, valutando la variabile riguardante la dimensione dei dati, si è sviluppato il modello mantenendo le altre due ferme ad una categoria e cambiando solamente la dimensione dei dati.

Dovendo quindi sviluppare un modello ad una sola variabile principale alla volta, sono state scelte per ognuna delle variabili marginali dei valori base da attribuirgli. La dimensione dei dati pari a 600 libri, la dimensione dello spazio vettoriale pari a 50 e infine la dimensione della finestra di contesto pari a 5. Questi valori sono frutto di alcune analisi esplorative svolte inizialmente per capire l’andamento della valutazione del modello e quindi ritenuti soddisfacenti.

4.6 Risultati e Interpretazione

I grafici sottostanti [Fig. 4.2] mostrano i risultati ottenuti durante la prima fase di valutazione delle tre variabili. Ogni grafico riporta sull’asse delle ordinate l’indice di correlazione misurato tra il giudizio del *query dataset* e quello ottenuto dal modello; sull’asse delle ascisse invece sono riportate le differenti categorie di ogni variabile.

Inoltre, ogni linea spezzata rappresenta un *query dataset* e il suo andamento in base alla categoria della variabile di interesse.

Nel primo grafico della figura 4.2, ossia la valutazione con variabile la dimensione dei dati, ovvero il numero di libri presi in considerazione, vediamo abbastanza chiaramente come un aumento della disponibilità di documenti di testo favorisca una maggiore capacità del nostro modello nel comprendere le forme di analogia e le similarità tra le parole; infatti, l'andamento del punteggio dell'indice di Pearson aumenta all'aumentare del numero di libri. A grandi linee, seppur un paio di *query dataset* presentano delle forme di outliars in certi casi, i restanti portano a pensare che un aumento della dimensione dei dati rafforzi la precisione nella valutazione da parte del modello in modo crescente e lineare.

Per quanto riguarda la dimensione dei vettori delle parole ottenuti nel *word embedding*, sembra non esserci un miglioramento, ma una sorta di stazionarietà nel comprendere il testo. È anche da notare che alcuni query-set portano ad un peggioramento di comprensione.

Infine, l'aumento della dimensione della finestra presenta una forma ambigua nella comprensione. Nel complesso non sembra esserci un miglioramento notevole, anche se alcuni *query dataset* abbaiano un andamento leggermente crescente, invece altri stazionario. Non si può quindi trarre una conclusione riguardo all'aumento o diminuzione di precisione.

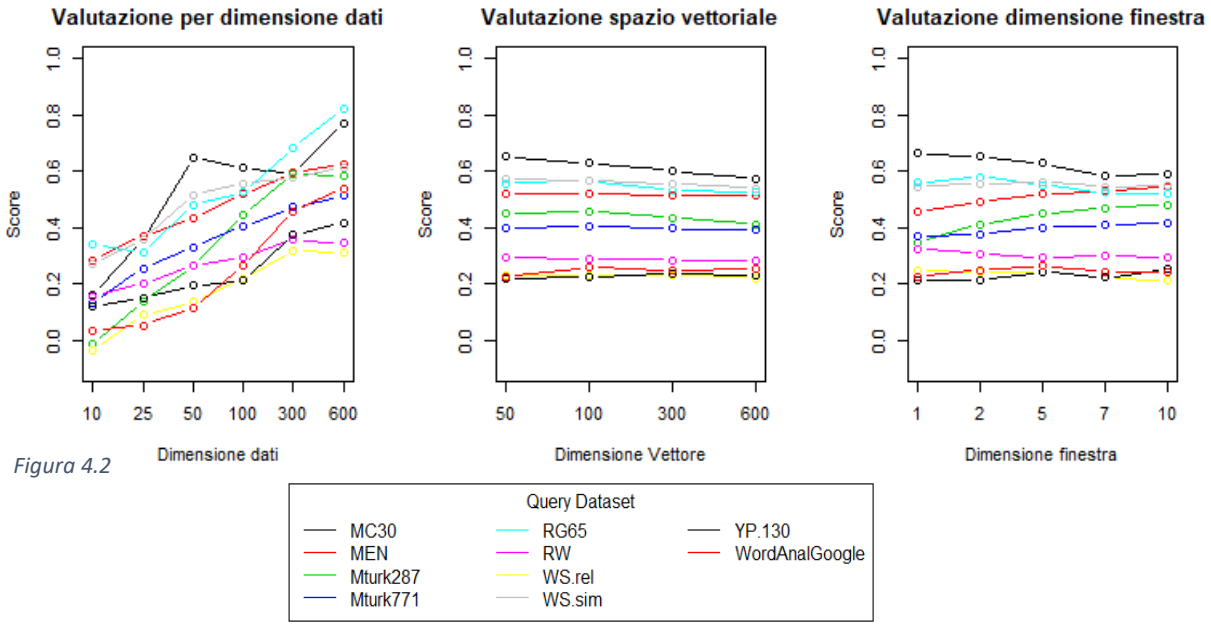

Figura 4.2

Un altro possibile strumento per visualizzare più in generale l'andamento dei modelli sotto valutazione può essere tramite l'ausilio dei box-plot [Fig. 4.3]. Questo strumento fornisce un riassunto della valutazione per ogni categoria di ogni variabile delle *query dataset* utilizzate. I grafici come in precedenza hanno lo stesso significato come valori degli assi, cambia solo il modo di visualizzazione. Un aspetto che si può cogliere in questa rappresentazione riguarda come i differenti *query dataset* si differenziano nella valutazione di ogni variabile, facendo pensare che ognuno di essi porta con sé differenti valutazioni semantiche fatte dalle persone.

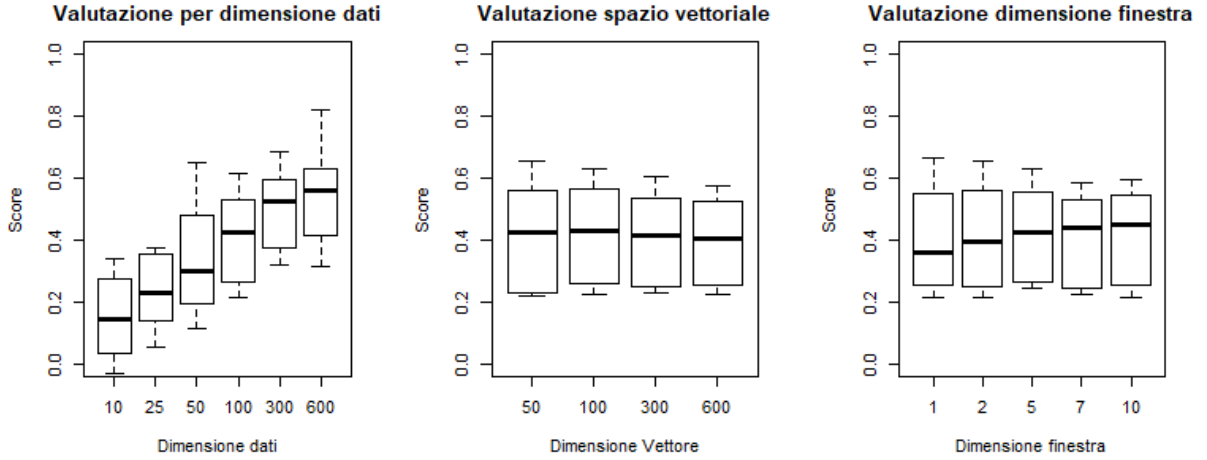

Figura 4.3

È importante sottolineare che l'uso del box-plot nella rappresentazione grafica può risultare limitata, data una bassa numerosità di *query dataset* per ogni categoria delle variabili. Però rimane comunque uno strumento utile per una comprensione generale.

Invece, i grafici della figura 4.2 ci permettono di effettuare una analisi visiva più specifica in base al *query dataset* di riferimento, potendo seguire l'andamento dei valori ottenuti nella valutazione per ogni variabile.

Essendo presente un'evidenza maggiore per l'importanza della dimensione dei dati sotto studio, nella tabella sottostante vengono riportati i valori della valutazione del modello in base alla grandezza dei dati e il rispettivo *query dataset* [Fig. 4.4].

	MC30		MEN		Mturk287		Mturk771		RG65		RW		WS.rel		WS.sim		YP.130		WordAnalGoogle	
10 Books	0.16	0.6	0.28	0.55	-0.01	0.44	0.13	0.6	0.34	0.51	0.16	0.09	-0.03	0.54	0.27	0.55	0.12	0.73	0.04	0.48
25 Books	0.36	0.77	0.37	0.66	0.14	0.57	0.26	0.74	0.31	0.74	0.2	0.19	0.09	0.69	0.36	0.68	0.15	0.76	0.05	0.63
50 Books	0.65	0.87	0.43	0.74	0.26	0.7	0.33	0.81	0.48	0.85	0.26	0.26	0.14	0.74	0.51	0.76	0.20	0.85	0.12	0.72
100 Books	0.61	1	0.52	0.81	0.45	0.78	0.40	0.87	0.53	1	0.30	0.35	0.22	0.8	0.56	0.81	0.21	0.88	0.26	0.82
300 Books	0.59	1	0.59	0.95	**0.59**	0.91	0.47	0.97	0.68	1	**0.36**	0.52	**0.32**	0.9	0.58	0.91	0.38	0.95	0.46	0.92
600 Books	**0.77**	1	**0.63**	0.98	0.58	0.92	**0.51**	0.97	**0.82**	1	0.34	0.61	0.31	0.9	**0.61**	0.92	**0.42**	0.95	**0.54**	0.93

Figura 4.4: Valutazione con ponderazione di frequenza

Per ogni *query dataset* vengono riportati due valori in corrispondenza del numero di libri utilizzati: il punteggio ottenuto e la proporzione di *query* utilizzate nella valutazione. Per il punteggio, inoltre, sono indicati in rosso i valori quando il test per valutare l'ipotesi nulla che l'indice di correlazione fosse pari a 0 fosse non rifiutato, quindi non valido nelle considerazioni. In grassetto viene evidenziato il valore con il punteggio più alto per variabile dimensione dati.

L'aspetto della proporzione del numero di query utilizzate durante la valutazione è importante, fornisce una indicazione di come il problema derivante dalla qualità delle stesse *query* prese in considerazione sia stato poi effettivamente affrontato. Maggiore sarà la proporzione, maggiormente potremo ritenerci parzialmente soddisfatte della valutazione, perché vuol dire che, seppur una discrepanza temporale causante di una possibile mutazione del linguaggio, l'adattabilità di alcuni query dataset di oggi non è del tutto da scartare. Inoltre, tramite un miglioramento della qualità tramite la ponderazione delle query, si rafforzerebbe di più questa idea del loro utilizzo.

Come ultima cosa paragoniamo i risultati ottenuti ai lavori [20] e [21] accennati precedentemente che hanno utilizzato il modello CBOW. [Fig. 4.5]

	MC30	MEN	Mturk287	Mturk771	RG65	RW	WS.rel	WS.sim	YP.130	WordAnalGoogle
[3]	NA	0.71	NA	NA	0.74	NA	0.57	0.72	NA	0.52
[4]	0.75	0.72	0.67	0.64	0.81	0.53	0.53	0.74	0.41	0.71

Figura 4.5: Valutazione modelli con architettura CBOW i lavori [20] e [21]

Vediamo che per alcune *query dataset* si hanno dei risultati abbastanza buoni, altre hanno ottenuto risultati comunque discreti. Possiamo ritenerci allo stesso modo soddisfatti.

4.7 Conclusioni finali di valutazione del modello scelto

Le considerazioni finali riguardanti la valutazione sono state già lievemente accennate in precedenza, ma ne sono sorte altre.

Ci stupisce che un possibile aumento della dimensione del vettore parola non abbia un impatto così notevole nella valutazione e miglioramento del nostro modello, o almeno nel nostro caso. Seppur comunque nella letteratura di questi lavori, e anche a livello logico, un aumento di questa variabile dovrebbe portare maggiore precisione nel *word embedding*, riteniamo che per questo lavoro non abbia influenza. Alcune cause, ipotetiche, possono essere dovute ad una scarsa dimensione dei dati utilizzati, oppure anche a possibili errori di sviluppo del modello, o ancora al tipo di dati che sono stato utilizzati. Non escludiamo quindi che in possibili lavori futuri si possa porre maggiore attenzione a riguardo.

La finestra di contesto ha portato risultati ambigui e qui ci sarebbe da fare una considerazione più specifica. Aumentare la dimensione della finestra nel *word embedding* vuol dire che l'interpretazione della parola sotto osservazione sarà influenzata anche da termini che saranno più distanti in quella frase, quindi, si direbbe che due parole abbiano un'influenza reciproca anche quando sono più lontane nelle frasi. Viceversa, se la dimensione della finestra diminuisce.

Per quanto riguarda la dimensione dei dati, si decide che quest'ultima variabile sia importante e da prendere in considerazione. Una possibile strategia per rafforzare questa idea sarebbe quella di effettuare un test statistico per valutarne l'effettiva significatività, ma ancora una volta, la bassa numerosità dei query dataset a disposizione ci porta a non poterne fare utilizzo. Restiamo fermi quindi con una valutazione maggiormente descrittiva.

Concludendo, si sceglierà il modello che prende in considerazione 600 libri, il vettore parola sarà pari a 50 e la dimensione della finestra pari a 10.

5. Analisi bias di genere

5.1 Introduzione analisi

In questo capitolo verrà effettuata un'analisi che ha l'obiettivo di andare ad individuare approssimativamente la possibile presenza di bias di genere nei documenti di testo. Si effettuerà un'analisi di tipo descrittiva, con l'idea di mostrare possibili metodi di partenza di studio.

Per farlo si utilizzeranno i romanzi letterali presi in considerazione fino ad ora, il modello ottenuto nella fase di valutazione e dei *query dataset*.

A differenza della parte di valutazione, i query dataset non avranno al loro interno dei giudizi già assegnati, ma toccherà al nostro modello ottenerli, dai quali poi verranno tratte le varie conclusioni.

5.2 Metodologie di analisi

Per analizzare la possibile presenza di bias di genere verranno utilizzati due metodi simili con alla base lo stesso principio, ossia la misurazione della vicinanza spaziale tra due termini tramite il metodo del coseno. Tale metodo è già stato descritto e utilizzato nella fase di valutazione del modello, quindi, verranno fatti solamente alcuni cenni di ripasso.

Presi due vettori di numeri reali con uguale dimensione, il metodo della similarità del coseno misura la vicinanza nello spazio vettoriale dei due vettori. Il valore è compreso nell'intervallo [-1, 1], dove se otteniamo 1 allora i due vettori occuperanno lo stesso spazio, invece -1 i vettori saranno opposti uno all'altro. Come fatto nella fase di valutazione, si effettuerà una trasformazione per avere un risultato compreso tra [0, 1].

Nel nostro caso i vettori, i quali rappresentano parole, più vicini saranno, più occuperanno lo stesso spazio semantico. Lo spazio vettoriale sarà costruito tramite il modello ottenuto in precedenza sui romanzi letterali.

L'operazione che si farà sarà quella di misurare la vicinanza, in termini di spazio vettoriale, tra pronomi e/o nomi che fanno riferimento al genere femminili e maschile e termini che idealmente non hanno in sé nessuna accezione precisa di genere, ma sono neutrali. Si osserverà infine se alcuni termini compaiono maggiormente in contesti dove il soggetto è una persona di genere femminile o maschile.

A. Il primo metodo si occuperà di prendere tre categorie di termini che fanno riferimento a mansioni di lavoro, descrizione dell'apparenza e tratti della personalità.

Dopodiché si prendono due liste di parole che fanno riferimento a nomi e pronomi di genere femminile e maschile del nostro spazio vettoriale creato dal modello.

Operativamente si misurerà la vicinanza tra tutti i termini di ogni categoria con le liste di nomi e pronomi di genere femminile e maschile. Più precisamente ogni termine riferito alla categoria verrà comparato con tutti i termini del genere di riferimento e verrà effettuata una media dei valori ottenuti, così da avere un indice di quanto quel termine sia vicino al genere femminile o maschile.

Infine, per ogni categoria si prenderanno i primi cinque valori di ogni genere in base al punteggio più alto, ossia a quanto si trovano maggiormente vicini nello spazio vettoriale.

Questa analisi non è di tipo comparativa, ossia non si vuole andare a paragonare, ad esempio, quale professione è più vicina ad un genere rispetto ad un altro, ma si intende stilare una sorta di classifica per capire quale termine di una certa categoria è maggiormente vicino ai rispettivi generi.

B. Il secondo metodo lavora attraverso una idea di cluster preconfezionato. Utilizzeremo sempre le stesse due liste che fanno riferimento a nomi e pronomi di genere e li confronteremo questa volta con dei gruppi di termini che sono stati formati in base a stereotipi già conosciuti. Ad esempio, due gruppi a confronto saranno parole che ruotano attorno al concetto di casa, quindi idealmente affiancate al genere femminile, e al concetto di lavoro, generalmente affiancato al genere maschile.

Per spiegarne meglio il funzionamento proponiamo un esempio: prendiamo i gruppi 'a' e 'b', casa e lavoro; il gruppo 'a' parte dal presupposto che siano presenti parole che vengono affiancate tradizionalmente al genere femminile e il gruppo 'b' rispettivamente al genere maschile. Ogni aggettivo del gruppo 'a' verrà messo a confronto con prima i pronomi di genere femminile e poi maschile, misurandone la vicinanza attraverso il metodo del coseno, e se successivamente si osserverà che la media dei valori del genere femminile è maggiore a quella del genere maschile, la parola rimarrà nella categoria di partenza, altrimenti verrà spostata nel gruppo 'b'.

Infine, si misurerà con quale proporzione i termini del gruppo ‘a’ siano rimasti ‘correttamente’ all’interno dell’insieme, penalizzando però la eventuale presenza dei termini del gruppo ‘b’.

I valori finali sono compresi in un intervallo [-1, 1], dove se otteniamo l’estremo sinistro vuol dire che nel gruppo ‘a’ sono presenti solamente tutti i termini del gruppo iniziale ‘b’ e viceversa, invece, se otteniamo il valore 1 entrambe le categorie sono rimaste invariate e confermano di conseguenza la presenza di stereotipi. Se si ottenesse il valore 0 allora teoricamente sarebbe presente una forma di equità di genere nelle categorie. Con questo meccanismo è da intendere che sia il gruppo ‘a’ e ‘b’ avranno lo stesso punteggio, quindi, il valore finale indicherà quanto quella categoria sotto osservazione sarà funzione di stereotipi di genere.

5.3 *Query dataset*

I *query dataset* utilizzati fanno riferimento ai lavori [31] e [32]. Sono stati leggermente modificati in base alla necessità di lavoro.

La lista dei pronomi e nomi dei generi sono i seguenti:

Female	Male
she	he
hers	his
her	him
woman	man
girl	boy
herself	himself
female	male
women	men
girls	boys
females	males

Per quanto riguarda i dati utilizzati nel primo metodo, abbiamo a disposizione tre categorie differenti con le rispettive numerosità: apparenza(25), occupazione(47) e personalità(423).

Invece nella seconda fase dove utilizziamo delle forme di cluster preconfezionate, abbiamo le seguenti coppie di categorie: buono/cattivo, arte/scienza e casa/lavoro [Fig. 5.1].

Arte	Art	dance	dancing	sing	singing	paint	painting	song	draw	drawing
Science	science	scientist	chemistry	physic	engineer	space	astronaut	chemical	microscope	math

Good	happiness	happy	fun	fantastic	lovable	magical	delight	joy	relaxing	honest	excited	laughter	lover	cheerful
Bad	torture	murder	abuse	wreck	die	disease	disaster	mourning	virus	killer	nightmare	stress	kill	death

Home	baby	house	home	wedding	kid	family	marry	clean
Work	work	office	job	business	economy	trade	money	finance

Figura 5.1

5.4 Risultati analisi

Il modello utilizzato per il *word embedding* sui nostri dati sembra confermare la presenza di termini più indirizzati ad un tipo di genere rispetto all'altro, molto legati agli stereotipi sociali [Fig. 5.2]. Ad esempio, per quanto riguarda termini riferiti alla personalità, la presenza di stereotipi è maggiormente marcata; il genere maschile è descritto più come una persona capace, forte e di successo, invece il genere femminile come un individuo con una personalità debole e fragile. Per la categoria che riguarda l'apparenza il genere femminile è connotato da aggettivi che riguardano sfumature di oggettificazione legate alla purezza e alla sessualità, invece per il genere maschile il ritratto sembra molto lontano dall'immagine femminile. Infine, la categoria sull'occupazione marca parzialmente i ruoli di genere; non c'è, o almeno sembra, una netta polarizzazione dei ruoli, forse anche dovuta dal fatto che i dati sotto studio sono poco comparabili con professioni come le conosciamo oggi.

Personalità	
Man	Woman
kind	amiable
capable	youthful
courageous	timid
Intelligent	charming
Proud	maternal

Apparenza	
Man	Woman
fat	voluptuous
strong	beautiful
healthy	attractive
ugly	handsome
beautiful	sensual

Professione	
Man	Woman
soldier	nurse
artist	dancer
teacher	artist
nurse	housekeeper
doctor	athlete

Figura 5.2

L'analisi riguardante gruppi preconfezionati sembra dare risultati simili a ciò che ci aspettavamo; per tutte e tre le categorie l'indice utilizzato ci dice che indicativamente i

termini sono rimasti all'interno del gruppo di riferimento allo stereotipo di genere, infatti i valori tendono verso 1 [Fig. 5.3].

Good/Bad	Home/Work	Art/Science
0,714	0,625	0,5

Figura 5.3

Osservando la figura 5.4, vediamo quali termini, in rosso, si sono spostati dal gruppo iniziale, quindi non identificandosi nel genere attribuitogli in partenza.

Home	baby	house	home	wedding	kid	family	marry	clean
Work	work	office	job	business	economy	trade	money	finance

Good	happiness	happy	fun	fantastic	lovable	magical	delight	joy	relaxing	honest	excited	laughter	lover	cheerful
Bad	torture	murder	abuse	wreck	die	disease	disaster	mourning	virus	killer	nightmare	stress	kill	death

Arte	Art	dance	dancing	sing	singing	paint	painting	song	draw	drawing
Science	science	scientist	chemistry	physic	engineer	space	astronaut	chemical	microscope	math

Figura 5.4

5.5 Conclusioni

Ovviamente le analisi fatte non sono rappresentative per trarre delle conclusioni definitive riguardanti la presenza di bias di genere nei documenti di testo studiati, ma possono essere utilizzati come una analisi descrittiva iniziale dà poi approfondire con metodi più efficaci.

Un grosso limite, ad esempio, dell'analisi attraverso i cluster riguarda la rigidità con la quale vengono assegnati i termini ad un gruppo in base alla loro posizione; è possibile che un termine occupi lo stesso spazio vettoriale per entrambi i generi, ma è sufficiente che sia leggermente più vicino ad uno che viene categorizzato in un modo; dunque, si potrebbe eventualmente valutare quanto diversamente distanti debbano essere i due generi dallo stesso termine per effettivamente indirizzare quest'ultimo in una categoria: test statistici appropriati possono essere una soluzione valida.

È molto importante sapere che quando si effettuano queste tipo di analisi con l'intenzione di trarre delle conclusioni, l'utilizzo di metodi inferenziali è fondamentale perché ci permette di dire quando due soggetti messi a confronto possono essere

ritenuti significativamente diversi, invece, le analisi descrittive sono uno strumento di partenza importante per capire un probabile andamento del fenomeno.

Un altro problema ricade nella scelta dei *query dataset* da utilizzare, sia sotto un punto di vista qualitativo che quantitativo; come è stato fatto nella analisi dei cluster, la numerosità utilizzata è irrisoria se si vogliono trarre delle conclusioni più consistenti; anche in questo caso la scelta è dovuta dall'intenzione di guardare sotto un aspetto di comprensione del metodo utilizzato. Anche per quanto riguarda la qualità della scelta delle *query* è necessario avere accortezza e precisione sotto un punto di vista linguistico.

6. Riflessioni conclusive

Lo studio appena effettuato ha mostrato i possibili metodi che possono essere utilizzati nei lavori di analisi dei bias nei documenti testo, con un riguardo maggiore ai bias di genere. Come visto durante il lavoro, trovare degli strumenti che siano perfetti e standardizzati da applicare a scatola chiusa non è possibile, ma è necessario un'attenzione maggiore.

L'elaborazione del linguaggio naturale da parte di una macchina è un processo che non può essere solo assegnato ad un unico modello o algoritmo, ma richiede che siano svolte delle fasi di sviluppo prima, durante e dopo il processo.

I modelli vanno interpretati e analizzati in modo meticoloso e attento; il rischio è di creare uno strumento che funzioni solamente al fine di ottenere i risultati desiderati, non potendo però generalizzarlo ad altri possibili casi dello stesso ambito. Le assunzioni iniziali sono fondamentali al fine di avere dei modelli che sia sufficiente informativi e per i quali ci si possa fidare, quindi avendo una forte struttura nello sviluppo.

Per quanto riguarda i dati sotto forma di documenti di testo bisogna avvalersi, anche qua, di importanti assunzioni derivanti dalla linguistica computazionale e la semantica distribuzionale, perché nel caso quest'ultime siano deboli, i modelli creati tramite *word embedding* assumerebbero altro significato. La linguistica in questi lavori non deve essere vista solamente come una cornice da utilizzare a piacimento, ma necessita di una buona dose di consapevolezza di come la si stia usando.

L'altro aspetto importante riguardava i bias di genere e la loro possibile presenza nei documenti di testo. Seppure i nostri modelli ci possano dare delle informazioni utili sulla semantica di un testo, abbiamo visto che al loro interno sono presenti alcune lacune sotto un punto di vista etico. Il funzionamento di un processo di machine learning necessita indubbiamente di essere analizzato non solo per quello che fa, ma occorre porre un'attenzione che riguarda le possibili conseguenze che quest'ultimo possa avere.

Come si è visto, infatti, il metodo utilizzato di *word embedding*, seppure funzionasse nella fase di valutazione generale, quando è stato poi applicato a possibili esempi di

distorsioni relativi a stereotipi di genere ha mostrato delle carenze sotto un punto di vista etico sociale.

Uno delle domande che forse bisognerebbe porsi quando si sviluppano questi strumenti è se applicandolo alla realtà ci possano essere delle possibili conseguenze che vadano a penalizzare o discriminare singoli o gruppi di individui; in tal caso sarebbe necessario cercare di capire dove possa essere una possibile falla e in che modo si possa prevenire.

Concludendo, i dati sui quali lavoriamo raccontano ciò che siamo e ciò che eravamo, riescono a cogliere qualsiasi aspetto e dettaglio, di conseguenza se il nostro modello imparerà da essi sarà successivamente intriso di distorsioni e pregiudizi. Tocca quindi a noi riuscire ad elaborare degli strumenti e metodi che siano una sorta di filtro dell'etica che dividano passato e futuro.

7. Bibliografia

[1] Batini, C. (2021). Enciclopedia dei dati digitali, Volume Primo: I dati sono una finestra sul mondo v1. Pubblicato con Licenza Creative Commons.

[2] Batini, C. (2022). Batini, C. (2022). Enciclopedia dei dati digitali, Volume terzo: L'etica dei dati digitali: l'Equità - Pubblicato con Licenza Creative Commons.. Milano : nessuno.

[3] https://en.wikipedia.org/wiki/Word_embedding

[4] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 446–457. DOI: https://doi.org/10.1145/3351095.3372843

[5] https://www.gutenberg.org/

[6] https://www.treccani.it/vocabolario/bias/

[7] Ricardo Baeza-Yates. 2018. Bias on the web. Commun. ACM 61, 6 (June 2018), 54–61. DOI: https://doi.org/10.1145/3209581

[8] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. DOI: https://doi.org/10.1145/3236386.3241340

[9] Milad Nasr and Michael Carl Tschantz. 2020. Bidding strategies with gender nondiscrimination constraints for online ad auctions. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 337–347. DOI:https://doi.org/10.1145/3351095.3375783

[10] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21). Association for Computing Machinery, New York, NY, USA, 249–254. DOI:https://doi.org/10.1145/3406522.3446033

[11] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21). Association for Computing Machinery, New York, NY, USA, 249–254. DOI:https://doi.org/10.1145/3406522.3446033

[12] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In

Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.

[13] https://www.treccani.it/vocabolario/sesso/

[14] https://www.treccani.it/vocabolario/gender/

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” CoRR, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[16] https://radimrehurek.com/gensim/models/word2vec.html

[17] Zellig S. Harris (1954) Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520

[18] Lenci, Alessandro and John Stockton Littell. “Distributional semantics in linguistic and cognitive research.” The Italian Journal of Linguistics 20 (2008): 1-32.

[19] A. Bakarov, “A survey of word embeddings evaluation methods,” CoRR, vol. abs/1801.09536, 2018. [Online]. Available: http://arxiv.org/abs/1801.09536

[20] Schnabel et al., 2015. Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *EMNLP*, pages 298–307.

[21] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang and C.-C. Jay Kuo (2019), "Evaluating word embedding models: methods and experimental results", APSIPA Transactions on Signal and Information Processing: Vol. 8: No. 1, e19. http://dx.doi.org/10.1017/ATSIP.2019.12

[22] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer, “Problems with evaluation of word embeddings using word similarity tasks,” arXiv preprint arXiv:1605.02276, 2016.

[23] G. A. Miller and W. G. Charles, “Contextual correlates of semantic similarity,” Language and cognitive processes, vol. 6, no. 1, pp. 1–28, 1991.

[24] E. Bruni, N.-K. Tran, and M. Baroni, “Multimodal distributional semantics,” Journal of Artificial Intelligence Research, vol. 49, pp. 1–47, 2014.

[25] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, “A word at a time: computing word relatedness using temporal semantic analysis,” in Proceedings of the 20th international conference on World wide web. ACM, 2011, pp. 337–346.

[26] G. Halawi, G. Dror, E. Gabrilovich, and Y. Koren, “Large-scale learning of word relatedness with constraints,” in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 1406–1414.

[27] H. Rubenstein and J. B. Goodenough, “Contextual correlates of synonymy,” Communications of the ACM, vol. 8, no. 10, pp. 627–633, 1965.

[28] T. Luong, R. Socher, and C. Manning, “Better word representations with recursive neural networks for morphology,” in Proceedings of the Seventeenth Conference on Computational Natural Language Learning,
2013, pp. 104–113.

[29] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pas¸ca, and A. Soroa, “A study on similarity and relatedness using distributional and wordnetbased approaches,” in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009, pp. 19–27.

[30] P. D. Turney, “Mining the web for synonyms: Pmi-ir versus lsa on toefl,” in European Conference on Machine Learning. Springer, 2001, pp. 491–502.

[31] https://github.com/nikhgarg/EmbeddingDynamicStereotypes/tree/master/data

[32] Charlesworth, Tessa E. S., Victor Yang, Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji. “Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words.” Psychological Science 32, no. 2 (February 2021): 218–40. https://doi.org/10.1177/0956797620963619.